SciDAC's Earth System Grid Center for Enabling Technologies

Semiannual Progress Report

October 1, 2010 through March 31, 2011

Principal Investigators

Dean N. Williams[3], Ian T. Foster[1], and Don E. Middleton[6]

The Earth System Grid Center for Enabling Technologies Team:

Rachana Ananthakrishnan[1], Neill Miller[1], Frank Siebenlist[1]
Mehmet Balman[2], Junmin Gu[2], Vijaya Natarajan[2], Arie Shoshani[2], Alex Sim[2],
Gavin Bell[3], Robert Drach[3], Michael Ganzberger[3],
Jim Ahrens[4], Phil Jones[4],
Daniel Crichton[5], Luca Cinquini[5],
David Brown[6], Danielle Harper[6], Nathan Hook[6], Eric Nienhouse[6], Gary Strand[6], Hannah Wilcox[6],
Nathaniel Wilhelmi[6], Stephan Zednik[6],
Steve Hankin[7], Roland Schweitzer[7],
Meili Chen[8], Ross Miller[8], Galen Shipman[8], Feiyi, Wang[8],
Peter Fox[9], Patrick West[9], Stephan Zednik[9],
Ann Chervenak[10], and Craig Ward[10]

Climate simulation data are now securely accessed, monitored, cataloged, transported, and distributed to the national and international climate research community.

[1] Argonne National Laboratory
[2] Lawrence Berkeley National Laboratory
[3] Lawrence Livermore National Laboratory
[4] Los Alamos National Laboratory
[5] National Aeronautics and Space Administration
[6] National Center for Atmospheric Research
[7] National Oceanic and Atmospheric Administration
[8] Oak Ridge National Laboratory
[9] Rensselaer Polytechnic Institute
[10] University of Southern California, Information Sciences Institute

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Auspices

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Contents

1 Executive Summary

This report summarizes work carried out by the Earth System Grid Center for Enabling Technologies (ESG-CET) from October 1, 2010 through March 31, 2011. It discusses ESG-CET highlights for the reporting period, overall progress, period goals, and collaborations, and lists papers and presentations. To learn more about our project and to find previous reports, please visit the ESG-CET Web sites: http://esg-pcmdi.llnl.gov/ and/or https://wiki.ucar.edu/display/esgcet/Home. This report will be forwarded to managers in the Department of Energy (DOE) Scientific Discovery through Advanced Computing (SciDAC) program and the Office of Biological and Environmental Research (OBER), as well as national and international collaborators and stakeholders (e.g., those involved in the Coupled Model Intercomparison Project, phase 5 (CMIP5) for the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5); the Community Earth System Model (CESM); the Climate Science Computational End Station (CCES); SciDAC II: A Scalable and Extensible Earth System Model for Climate Change Science; the North American Regional Climate Change Assessment Program (NARCCAP); the Atmospheric Radiation Measurement (ARM) program; the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA)), and also to researchers working on a variety of other climate model and observation evaluation activities.

The ESG-CET executive committee consists of Dean N. Williams, Lawrence Livermore National Laboratory (LLNL); Ian Foster, Argonne National Laboratory (ANL); and Don Middleton, National Center for Atmospheric Research (NCAR). The ESG-CET team is a group of researchers and scientists with diverse domain knowledge, whose home institutions include eight laboratories and two universities: ANL, Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), LLNL, NASA/Jet Propulsion Laboratory (JPL), NCAR, Oak Ridge National Laboratory (ORNL), Pacific Marine Environmental Laboratory (PMEL)/NOAA, Rensselaer Polytechnic Institute (RPI), and University of Southern California, Information Sciences Institute (USC/ISI). All ESG-CET work is accomplished under DOE open-source guidelines and in close collaboration with the project's stakeholders, domain researchers, and scientists.

Through the ESG project, the ESG-CET team has developed and delivered a production environment for climate data from multiple climate model sources (e.g., CMIP (IPCC), CESM, ocean model data (e.g., Parallel Ocean Program), observation data (e.g., Atmospheric Infrared Sounder, Microwave Limb Sounder), and analysis and visualization tools) that serves a worldwide climate research community. Data holdings are distributed across multiple sites including LANL, LBNL, LLNL, NCAR, and ORNL as well as unfunded partners sites such as the Australian National University (ANU) National Computational Infrastructure (NCI), the British Atmospheric Data Center (BADC), the Geophysical Fluid Dynamics Laboratory/NOAA, the Max Planck Institute for Meteorology (MPI-M), the German Climate Computing Centre (DKRZ), and NASA/JPL.

As we transition from development activities to production and operations, the ESG-CET team is tasked with making data available to all users who want to understand it, process it, extract value from it, visualize it, and/or communicate it to others. This ongoing effort is extremely large and complex, but it will be incredibly valuable for building "science gateways" to critical climate resources (such as CESM, CMIP5, ARM, NARCCAP, Atmospheric Infrared Sounder (AIRS), etc.) for processing the next IPCC assessment report. Continued ESG progress will result in a production-scale system that will empower scientists to attempt new and exciting data exchanges, which could ultimately lead to breakthrough climate science discoveries.

2 Reporting Period Highlights

In support of the IPCC AR5, the Earth System Grid Federation (ESGF) is building on the success of ESG to resolve large-scale data and management issues facing CMIP5. CMIP5 drives all ESGF work and is the primary motivator for collaborators to add other climate data sets, including observational and reanalysis data sets to the ESGF effort.

The early arrival of ESG data sets was a major highlight during this reporting period. As of March 31, 2011, the federated repository has collected and disseminated INMCM4 (Russian pre-industrial control model output) and HadGEM2-ES (United Kingdom (UK) Met Office Hadley Centre atmospheric 6, hourly model output). More importantly, ESGF is receiving reports of scientific results that have been achieved through use of the long-awaited CMIP5 model archive. INMCM4 data are being served from the Program for Climate Model Diagnosis and Intercomparison (PCMDI) model data node, and HadGEM2-ES output is being served from the BADC data node. These data sets can be viewed from any of the eight ESGF distributed gateways, which deliver seamless access to the federated archive. At the moment, data exploration is available only via Web browsers.

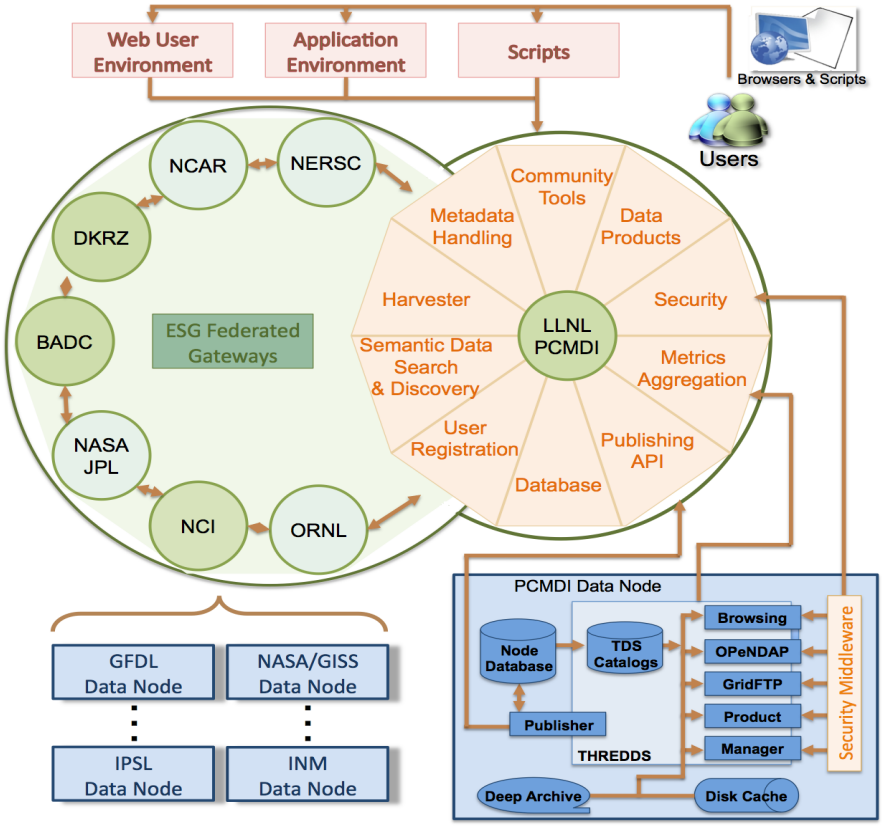

Figure 1: The figure shows how users can access ESGF data using Web browsers, script, and in the near future, client applications. ESGF is separated into two parts: ESGF gateways (indicated in green), and ESGF data nodes (indicated in blue). Gateways handle user registration and management and allow users to search, discover, and request data. Data nodes are located where the data resides, allowing data to be published (or exposed) on disk or through tertiary mass store (i.e., tape archive) to any gateway. In addition, data nodes handle data reduction, analysis, and visualization. ESGF currently comprises eight national and international gateways, four of which hold special status in housing CMIP5/AR5 replication data sets: LLNL/PCMDI, BADC, DKRZ, and the Australian National University National Computational Infrastructure (NCI). Users have access to all data from the federation regardless of which gateway is used.

For CMIP5, ESGF will archive 3.3 petabytes (PB, 1 PB = 10^{15} bytes) of officially requested data. About half of the data (e.g., data on surface temperature and precipitation), will likely be in high demand by researchers, so the five ESGF partners will replicate this data. The modeling groups "manually" control access to the data until the data providers fill out the metadata questionnaire (developed by our European partners). This questionnaire is completed before the metadata has passed quality control level 1. Once providers have published their data and completed the questionnaire, the data will automatically be made accessible to everyone and replicated to the other sites.

2.1.1 ESG Release Schedule

Table 1: ESG source code release schedule.

1.3.0—CMIP5 security and integration	**Goal:** ESG security updates and product service integration	**Date:** Monday, April 4, 2011

Gateway
• Spring Framework 3.0 upgrade.
• Security enhancements:
  • Centralized gateway registry.
  • Centralized white-list management.
• Gateway collation of data node manager metrics:
• Enhanced CMIP5 XML harvesting from METAFOR questionnaire atom feed.
• Demonstrate base Live Access Server (LAS) Product Service integration:
  • Enable basic subsetting services.
  • Enable basic data visualization services.
• New data set notification support.		

Data Node
• Include ESG Publisher graphical user interface (GUI) in the release of the data node:
  • Modified and added new features as requested by ESG staff and management.
  • ESG publisher provides a GUI interface, which has been updated to support data set versioning.
• Access metrics service in place for metrics fetching.
• New release of installation script.
• ESG publisher generates all metadata as prescribed by the Data Reference Syntax (DRS):
  • Data set versioning is supported as defined by DRS.
  • Support is enabled in the ESG publisher for retrieval of data sets using a DRS-defined syntax.
  • DRS controlled vocabulary is maintained within the ESG publisher configuration.
  • ESG publisher supports DRS filename conventions.
• ESG publisher maintains CMOR (Climate Model Output Rewriter) table information.
• ESG publisher interfaces with the ESG node installation script.
• ESG query services:
  • ESG data node suite connects with all ESG programming interfaces.
  • ESG publisher allows metadata queries to be directed to any ESG gateway.
• Support for other (non-CMIP) projects:
  • A data handler has been written for observational data associated with CMIP5.
  • A data handler is available for the Transpose Atmospheric Model Intercomparison Project (AMIP) project.		

1.4.0—CMIP5 distributed archive	**Goal:** ESG product services and user interface	**Date:** Monday, May 16, 2011

Gateway

- Data access by DRS Uniform Resource Locators (URLs).
- Advanced support for replica discovery and download:
  - Search of replicated data sets.
  - WGet script generation at remote gateway.
- Present data set version history in user interface (UI).
- Fully support deletion of data sets:
  - Verify database integrity after hard deletion.
  - Support logical deletion.
- Enhance data download workflows:
  - Improved script generation (via Data Cart capabilities, add WGet help, version requests).
  - Improved large file set selection.
- LAS product server integration (data sets and variable selection with auth-z):
  - Enable basic subsetting services.
  - Enable basic data visualization services.
- Manage file access locations using publisher:
  - Support for file rename, move, and remove use cases.
- New data set version notification support.
- Optional token support discontinued based on federation feedback.

Data Node

- Data node documentation and help instructions.
- Support for GridFTP services (GridFTP is the Globus high-performance FTP server):
  - ESG publisher allows configuration and generation of GridFTP access to URLs.
- Continue upgrade of installation script with new component integration.
- LAS "Product Server" integration:
  - ESG publisher generates LAS URLs for publication, which enables LAS access to all published data sets.
- Enhance ESG publisher GUI:
  - Design and build database metadata validation tool.
  - Incorporate MyProxy into the publisher GUI.
  - Add gateway query capability to the publisher GUI.
  - Add ESG list files to publisher GUI to show what files comprise a data set.
- Continue support for other (non-CMIP) projects:
  - Data handler development.
  - Product tool integration.
- DRS refinement and CMIP5 data conventions updates (as needed).
- Continue operational status of CMIP5.

1.5.0—CMIP5 – Baseline 3	**Goal:** Based on 1.4 system reviews and user feedback	**Date:** Monday, July 18, 2011

Gateway

Release schedule will be based on user feedback on the 1.4 release

Data Node

Release schedule will be based on user feedback on the 1.4 release

2.1.2 LLNL Readies for Implementation of CMIP5 (IPCC AR5) Data

PCMDI has strongly focused on the development, installation, and support of ESG data nodes to assist modeling centers with processing and integrating their data into the ESGF archive. Besides code development, this effort includes answering script installation and difficult technical hardware and network connection questions, working with DKRZ's quality control (QC) software, and modifying the Climate Model Output Rewriter (CMOR) for ESG metadata discovery. The ESG node is an amalgam of many disparate "best of breed" software products that are installed and configured together. PCMDI is in charge of writing and maintaining the software archive, repository, and installation code for the data node. The installation process should be as simple as possible for the end user (e.g., modeling centers or disparate data providers). To this end, the installer combines analogous set into a single installation unit.

Providing support to the wide-range of ESGF users, which includes modeling centers, data providers, scientist, and non-scientists, is of the utmost importance. Support efforts include bug tracking and fixing

and trouble shooting. As expected, it also includes interfacing with national and international users to address their specific system requirements. The LLNL team is also leading the development of several software products used for integrating and supporting the ESG distributed data platform, which includes the data node publisher, the metrics collection and distribution modules, the network and system monitoring modules, and user notification module. These software products must fit into the overall design of current and future ESG systems.

CMIP5 data is subjected to three levels of QC checking. Level 2, "conformance checks and subjective quality control," involves extensive examination of metadata and data for self-consistency and conformance to standards. These quality checks are performed using the "QC Tool" developed at DKRZ. For example, a variable must have a recognized "standard_name" attribute and its data values must lie within a range of values for that standard_name. Another example is that a file name must conform to the Data Reference Syntax (DRS) file naming specification, and the time range incorporated in the file name must be consistent with the range of values of the time variable in the file. DKRZ maintains a database of the results for every QC check performed anywhere in the world with their primary tool. Over the last few months, PCMDI has contributed to further development of the QC tool by testing and debugging earlier versions.

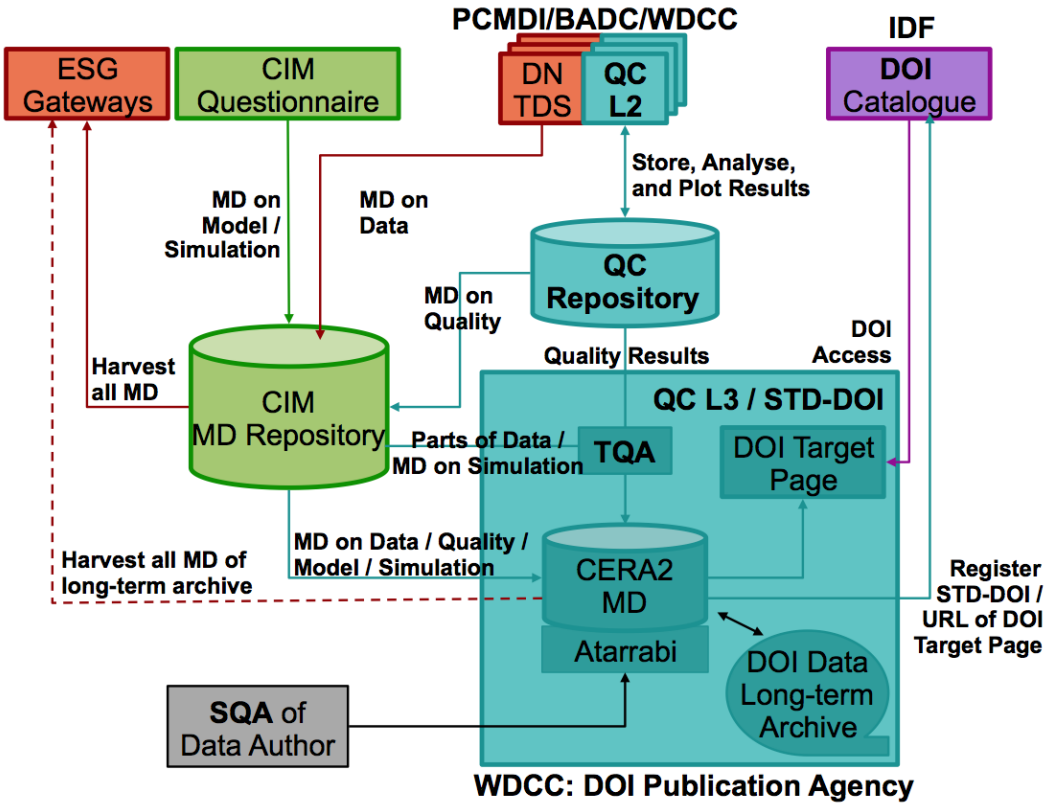

Figure 2: This diagram shows the CMIP5 QC process within the ESG environment. Acronyms are defined as follows: MD—Metadata, DN—Data Node, TDS—THREDDS Data Server, TQA—Technical Quality Assurance, SQA—Scientific Quality Assurance, QC—Quality Control, DOI—Digital Object Identifier, IDF—International DOI Foundation Member, WDCC—World Data Center for Climate, CIM—Common Information Model, STD-DOI—Publication and Citation of Scientific Primary Data (a project funded by the German Science Foundation). The QC service layer will be installed at three sites (PCMDI, BADC, and DKRZ). DOIs are issued after QC Level 3 has been confirmed. (See the previous (April 1, 2010 through September 20, 2010) progress report for the three stages of the QC data workflow.) (Image courtesy of Martina Stockhause from the DKRZ.)

For ESGF to discover all the metadata in its files and allow for advanced searching and server-side analysis, the data should follow Climate and Forecast (CF) conventions. For the CMIP5 project, in particular, PCMDI developed CMOR, which transforms climate simulation data into netCDF[1]/CF compliant files (CMOR is based on CF but also includes extra metadata specification for ESG). CMOR was completed and released to the community, however, updates to the software is an ongoing process. The current version, v2.5.7, contains "patches" to fix typos for improved error messaging. CMOR v3.0 is under development. This new version will include support for observational and reanalysis data, contain a possible CMOR checker, handle Gridspec files, and use the new LibCF library.

The International DOI Foundation (IDF) is an open membership consortium that includes both commercial and noncommercial partners. IDF developed and manages the Digital Object Identifier (DOI) system (www.doi.org) used in ESGF. The DataCite Consortium is a virtual registration agency for data publications within the DOI system, and includes "real" national registration agencies such as the German National Library of Science and Technology (http://www.tib.uni-hannover.de/en.html) for World Data Center for Climate (WDCC). Thus, the DOI hierarchy is as follows:

- IDF—the DOI system.
- DataCite—member of DOI system functioning as a registration agency for IDF.
- The German National Library of Science and Technology and other national registration agencies—member of international DataCite consortium; serves as a registration agency for national publication agencies such as WDCC.
- WDCC and other publication agencies—contract partner of national registration agencies; grants persistence of data and metadata related to DOI; DOIs are registered at DataCite (datacite.org and api.datacite.org).

2.1.3 NASA ESG Gateway Research & Development Highlights

Several ESGF partners (PCMDI, NASA, ORNL, NOAA, BADC, Institut Pierre Simon Laplace des Sciences de l'Environnement Global (IPSL), and others) have collaborated to define a specification for data formats and metadata conventions that can be adopted by all providers of observational data submitted to the ESGF global archive. This effort is taking shape with NASA/JPL data coming online in April and DOE's Climate Modeling Best Estimate (CMBE) data scheduled for uploading shortly thereafter. The ESGF collaboration has advanced in defining the next generation peer-to-peer architecture for a global infrastructure to support climate research. Notable progress has been made in the areas of security, federation registry, and support for client analysis. The JPL team has completed deploying an operational system for distributing NASA observations within the ESGF enterprise system.

2.1.4 ORNL ESG Gateway Portal Research & Development Highlights

The ORNL ESG-CET team has worked closely with the entire ESG-CET team to deploy the ESG v1.2 to the climate science community. The ORNL gateway is now serving data sets from the ARM program, the Carbon-Land Model Intercomparison Project (C-LAMP), the Carbon Dioxide Information Analysis Center (CDIAC), the Community Climate System Model (CCSM), and the Ultra High Resolution

[1] NetCDF (network Common Data Form) is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.

Global Climate Simulation (UltraHighRes), as well as observation data sets specifically developed in support of CMIP5. The ORNL team has transitioned most of the ESG existing data sets to the next-generation ESG infrastructure and is in the process of retiring the legacy infrastructure. ORNL has deployed ESG's next-generation data publication process, allowing ESG to capture full data set metadata for off-line data sets. Given this progress, the ESG infrastructure at ORNL is well poised to meet the needs of the climate science community.

2.1.5 LBNL Berkeley Storage Manager

Changes in requirements and suggestions from end users of the Bulk Data Mover (BDM) and DataMover-Lite (DML) have been integrated into the releases. DML now supports multiple ESG federation gateways and http downloads, and is fully integrated with the ESG authorization model and plug-ins. The DML web-start and stand-alone versions are available on the Web and have been used in production. BDM also includes new features for improved efficiency. BDM and ESG networking capabilities on ESGF have been tested extensively for ESG data set replication (e.g., Figure 1). LBNL/National Energy Research Scientific Computing Center (NERSC) gateway and data node have been deployed in production, serving IPCC AR4 CMIP3 data sets.

2.1.6 PMEL Product Delivery Services Highlights

The PMEL team accomplished several milestones with regard to the product services user interface. We integrated the LAS User Interface Client with the ESG security infrastructure. PMEL also improved the Ajax communication interfaces to facilitate cleaner and more user-friendly data discovery via the ESG gateway. Additionally, the team completed prototypes for including an ensemble axis in the comparison interface. The PMEL team also led a cooperative effort to investigate direct client access to the Open-source Project for a Network Data Access Protocol (OPeNDAP) servers protected by the ESG security infrastructure. Highlights from PMEL also include working with the ESG team to improve the publication process with regard to the automatic production product service metadata and configuration information during publication. The team also made many improvements to the Ferret-THREDDS Data Server (F-TDS) interface that provides on-the-fly analysis capabilities in ESG. The improvements include more complete metadata on analysis variables and the ability to serve more variable types (including character variables) through F-TDS. Finally, stack in cooperation with LLNL, PMEL has continued to test and refine the installation process for the ESG data node software.

2.1.7 ANL Security, Data, and Services Highlights

ANL has contributed to ESG's progress as described in Section 3, in particular, with the Globus Online integration and federation metadata management. The Globus Online integration facilitates leveraging solutions for data transfer and management, thus focusing on climate specific infrastructure and integration pieces. Integrating the metadata-based search capabilities allows user's to use the service from within ESG context, providing a seamless user experience, while using a data transfer solution that provides both high performance and reliability. Federation metadata management is key to defining and binding the ESG federation—system to define, and manage the metadata, and provision them across the ESGF nodes provides a collaborative experience for the end user. Metadata includes the trustroot configuration, various data and security services, and node information.

2.1.8 USC/ISI Replication Highlights

The data replication team at ISI deployed the ESG replication client to BADC and DKRZ for testing. The issues identified by expanding testing were analyzed and solutions were added to the software. In addition, the Globus Online service was added as an option for data transfers.

The replication client integrates functionality from several ESG providers. It queries the metadata catalog at the gateway; pulls the THREDDS catalog from the original publication site's data node; creates control files for either the BDM client or the GridFTP globus-url-copy client to transfer data; and invokes the publication API from LLNL to scan the replicated files, generate a Thematic Real-time Environmental Distributed Data Services (THREDDS) catalog for those files, and publish the mirrored data set to the gateway. The replication client compares the THREDDS catalog with files at the original publication site. The data set is only published if this comparison verifies the correctness of the THREDDS catalog.

2.1.9 RPI Highlights

RPI continues to make progress using OPeNDAP Hyrax as a drop-in replacement for the THREDDS Data Server (TDS). This work is being done in two parts. First, RPI is reconfiguring OPeNDAP Hyrax to read the THREDDS catalogs within the ESG data node. Second, they are adapting OPeNDAP Hyrax to communicate with a ferret server so that it generates different responses (e.g., Data Access Protocol (DAP) responses, data responses) within the ESGF enterprise system. The RPI team has spent time writing up requirements for both of these efforts, gathering information from different collaborators who currently work with ESG's Product Server, LAS, and TDS. The team has also contributed system diagrams and state diagrams for RPI.

Diagrams and requirements for the THREDDS catalog integration can be found at: http://tw.rpi.edu/web/project/ESG-CET/workinggroups/aggregation. OPeNDAP Hyrax can already read THREDDS catalogs, except for the NetCDF Markup Language (NcML) elements within the catalogs. Work is currently underway to remedy this issue. Diagrams and requirements for the Ferret integration can be found at: http://tw.rpi.edu/web/project/ESG-CET/workinggroups/thredds. With the modular framework provided by the OPeNDAP Back-End Server (BES), Ferret will provide ESG users with additional capabilities.

General improvements are being made to the OPeNDAP BES to create better installation services, administrative services, and modifications geared toward meeting the high-performance and scalability requirements of ESGF. RPI continues to work with collaborators within ESG on installing the ESG data node to test these requirements, and to create use cases and requirements for these activities. Use cases are needed to show how the ESG Product Server and OPeNDAP Hyrax will be integrated together. Recent discussions have lead to new requirements. Information on RPI's ongoing work for ESG-CET can be found at: http://tw.rpi.edu/web/project/ESG-CET.

RPI currently has the latest VMWare CentOS5 version of the data node installed on a newly purchased machine for developing ESG software-related products.

3 Component Progress

During this reporting period, the ESG-CET team has made progress toward meeting ESG-CET objectives, goals, and milestones. This section describes that work in greater technical detail and presents the components needed for the CMIP5 baseline release.

3.1 Gateway Software Development

The ESG gateway was extensively developed during this reporting period. Accomplishments included the major release of v1.2.0 and three milestone releases for v1.3.0. The ESG-CET team continued to refine the software engineering process, moving from an ad hoc process to a canonical Agile process later in this reporting cycle.

ESG Gateway v1.2.0: As documented in the last ESG-CET semiannual report, the team's plan was to release the critical CMIP5 Baseline Release, v1.2.0, in October of 2010 and move into production mode. Although a release candidate was issued in October, it took far longer to reach production mode than previously planned. One of the goals was to stress test the system with substantial volumes of real CMIP5 data, which was expected to arrive in October. Unfortunately, the data was not available at that time, and the team is currently still waiting on the modeling groups. The team also had to complete substantial international federation testing, which took time to ramp up with activities extending into the holidays. The team finally released v1.2.0 on February 11, 2011, and promptly moved into production mode. Version 1.2.0 was a major release, and included over 75 new features and bug fixes. Several updates were as follows:

- Completed baseline release for CMIP5 data management.
- Implemented federation-wide X509 certificate and OpenID-based authorization.
- Captured full CMIP5 DRS metadata.
- Automated Open Archives Initiative (OAI)-based federation metadata exchange using Resource Description Framework (RDF).
- Completed basic METAFOR CIM ingestion (collaboration with NOAA/ Global Interoperability Program (GIP)/Curator and METAFOR).
- Streamlined gateway installation and configuration.
- Integrated MyProxyLogon WebStart application into data download user interface.
- Captured missing DRS attributes during publication.

NCAR has been running v1.2.0 as its primary production service since February, and the system has performed well and reliably.

ESG Gateway Version 1.3.0: Work on v1.3.0 started in November 2010, and progressed through the international deployment and testing process of v1.2.0. To mitigate delays that are inherent in a large, international collaboration, NCAR transitioned to a full Agile development process, where units of work were broken into "sprints." As a result, milestone releases were achieved that were shared with the community for evaluation and testing.

On February 28, 2011, the ESG-CET team released v1.3.0 M1 (Milestone 1). This release included several important new features along with bug fixes. These upgrades included:

- Upgraded Spring 3.0 Framework.
- Upgraded Spring Security.
- Improved MyProxy service.
- Fixed data access and authorization workflow.

On March 15, 2011, the ESG-CET released v1.3.0 M2, which:

- Provided a centralized gateway registry for federation support.
- Improved user registration workflow.

On March 23, 2011 we released 1.3.0 M3:

- LAS product server integration.
- Metrics and notification support.
- Present data set version history in UI.

For greater details on the gateway source code release schedule, please see table 1.

The ESG-CET team plans to release v1.3.0 to the community in early April 2011, shortly after this reporting period. Version 1.3.0 will serve as the vehicle for releasing the gateway as an open-source development effort. The team has completed an analysis of the legal characteristics of hundreds of libraries required by the gateway, and has resolved virtually all of the related issues. Concurrently, a Contributor License Agreement (CLA) was developed. This agreement is a simplified version of the one used successfully by the Apache community. The ESG-CET CLA is aimed at ensuring the gateway can continue to serve as an open-source resource, while protecting the intellectual property of contributing entities. BADC, ANL, and ORNL have all executed CLA's and are engaged in development work.

BADC is developing a next-generation METAFOR CIM harvester that will provide sophisticated metadata ingest capability for the ESG gateway. ANL also has their own development branch of the gateway, and is integrating the Globus Online cloud-based data transfer capabilities as a new approach to managing and transferring large volumes of scientific data in the ESG environment.

In parallel with all the above-mentioned development activity, the ESG-CET team has also been engaged in activities aimed at integrating the LAS Confluence Server with the ESG gateway, and progress has been quite good. The team can now demonstrate basic functionality for a gateway+LAS configuration, and plans to release this configuration into production later this year.

Metrics and Summary: All of the work described in Section 3 of this document reflects extensive collaboration among core ESG-funded partners, the NOAA GIP effort, Curator, the European Union METAFOR project, and the international Global Organization for Earth System Science Portals (GO-ESSP) community, especially BADC and DKRZ.

The figure below depicts the cumulative data download volume for the NCAR ESG gateway, along with the last year of monthly data download volume.

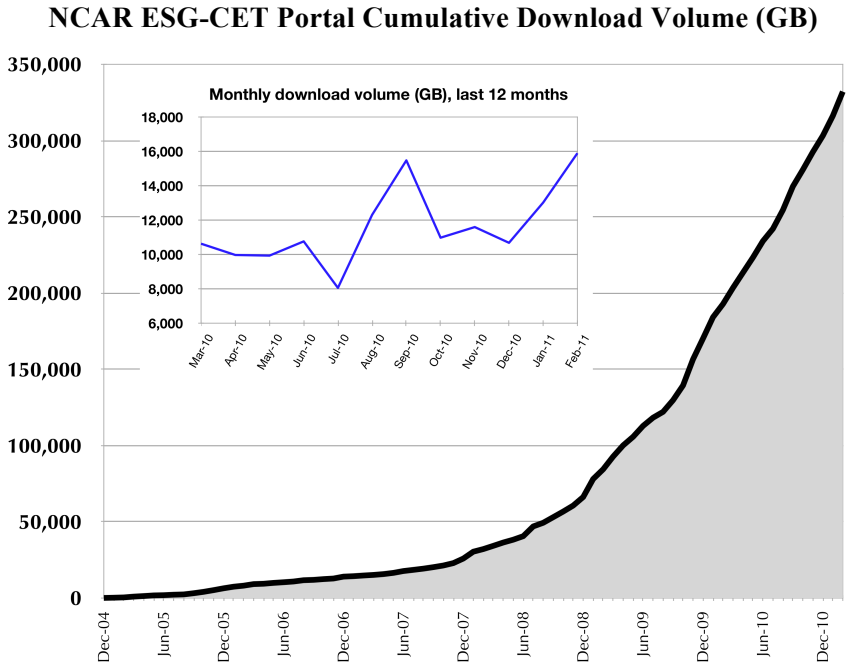

Figure 3: This graph provides additional download metrics. Volume is measured in gigabytes (GB).

3.2 Data Node Software Development

Derived Observation Data Product: CMBE observational data from the ARM program was specifically tailored to the atmospheric modeling community. The data are a condensed and integrated subset from the ARM data collection and include measurements that have undergone stringent QC processes. These measurements are usually best estimates or derived from several instruments and/or analysis data. Therefore, this observational data set is very well-suited for supporting IPCC CMIP5 model output simulation validation.

The CMBE observational data was rewritten in close accordance with the CMIP5 model data output specifications and requirements to improve transparency (in analysis and comparison) between the two data sets. This effort included:

1. Defining the appropriate standardized observational metadata (international collaboration of observational data centers).
2. Rewriting the data to meet stringent CMIP5 data requirements (using CMOR software to accomplish one variable per file data sets, CF-compliant NetCDF form, strict CMIP5 definition of appropriate variables).
3. Defining and verifying the ESG publication process (involved cooperation and coordination with ORNL and PCMDI teams to develop software and processes enabling the publication of observational data, and the observational metadata to be harvested and searchable in the ESG).
4. Testing publication of CMBE data in both the PCMDI and ORNL gateways. This process revealed problems how the gateway harvests metadata that is specific to the observational data; this issue is being currently resolved.

5. Continuing discussions within the observation data community regarding observational metadata standards. Several issues should be resolved very soon.

Data Format and Metadata Conventions for CMIP5 Observations: Starting in Fall 2010, JPL began collaborating closely with PCMDI to gather support across agencies for adopting common data and metadata conventions to structure observational data sets submitted to the CMIP5 archive. The intended goal is to facilitate using observations from different sources (satellites, station data, trajectories, and profiles) to evaluate and diagnose the model predictions. To accomplish this goal, the observational data sets must closely resemble the model output.

Two very successful NASA-sponsored workshops were held in September and November 2010 to address data format and conventions. These workshops were followed by intense collaborations between PCMDI, several NASA centers, ORNL, NOAA, BADC and IPSL. This multiagency, international effort resulted in specifying the internal structure of the observational data, the name and type of the embedded global attributes that will be harvested for search and discovery purposes, and the directory structure and filenames that must be adopted to provide homogeneity across data providers (CMIP5 Observation (OBS)). Currently, the community is defining a controlled vocabulary to accompany the metadata conventions, while at the same time configuring, and possibly modifying, the Climate Model Output Rewriter 2 (CMOR-2) program—originally developed to post-process model output—to suit the observation data. The first CMIP5 observational data sets from NASA (i.e., AIRS—satellite observation data) and ORNL (i.e., ARM—in situ station data, and CDIAC—in situ observation data) are expected to be published into the ESGF in Spring 2011.

For CMIP5 observation data set requirements, visit the following URL:

https://oodt.jpl.nasa.gov/wiki/display/CLIMATE/Data+and+Metadata+Requirements+for+CMIP5+Observational+Datasets

Description of Changes to PYTHON Publisher GUI: Figure 4 shows a sample screenshot of the Python Publisher GUI. Versioning is now maintained and displayed with its own column (see figure 4). In addition, new features include publishing and un-publishing, getting help via the Web browser, and viewing collected and displayed comments. In this example, the “Delete” option allows a user to remove data sets from the THREDDS server and gateway, while letting them keep it on the local database.

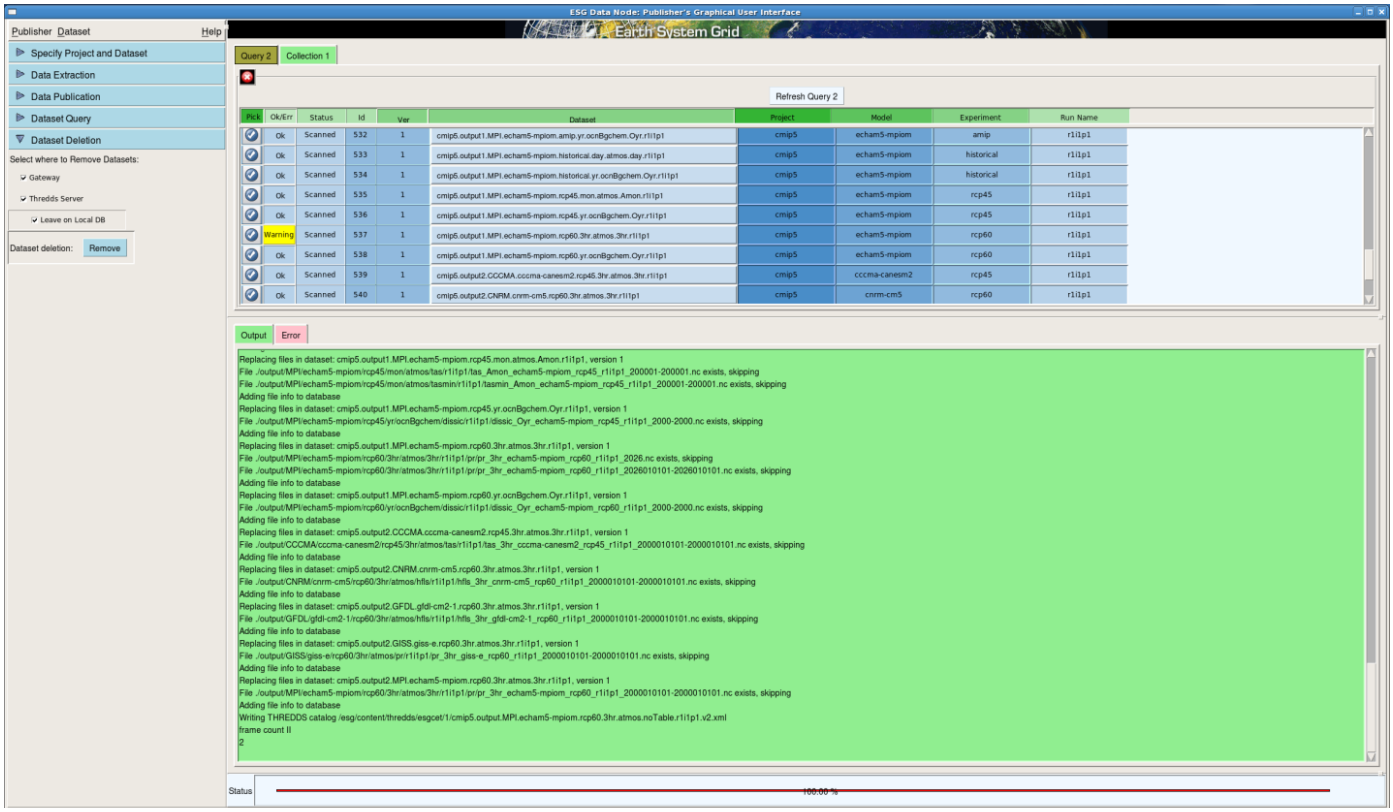

Figure 4: This is a screenshot of the current Python Publisher GUI based on Tkinter. The data node development team continued to add new features to the GUI for the current release of ESGF (v1.3.0).

Description of Changes to Java Publisher GUI: After reexamining the initial Java prototype GUI code, the ESG-CET team determined the code needed to be re-factored to move it into a model-view-controller (MVC) pattern. This conversion will ensure ESG-CET code developers will be able to change parts of the GUI as the needs of the community continue to evolve. This upgrade will also make the GUI easier to understand and maintain. The new Java-based Publisher GUI resides in the repository under the directory esg/node/publisher/gui. The GUI is designed with clear class delineation of responsibility; hence, new folders have been created for actions, dialogs, network, subcomponents, and utilities.

Building a GUI that consolidates responsibilities into one class can lead to a large, fragile, monolithic-looking code. As the ESG-CET team begins to add functionality and more classes to the GUI, especially to the model and controller aspects, it anticipates the need for more re-factoring. One change implemented was replacing the layout call GridBagLayout with SpringLayout. SpringLayout is a more flexible layout manager. It allows users to specify precise relationships between the edges of components under SpringLayout control, making it much simpler to use and maintain. The big advantage of this Java-based GUI over the Python-based GUI is its ability to start-up from a Web browser with remote access, meaning that the GUI does not need to be collocated on the same machine as the data. As a result, the data provider with the right credentials can initiate a remote GUI publisher and publish data anywhere.

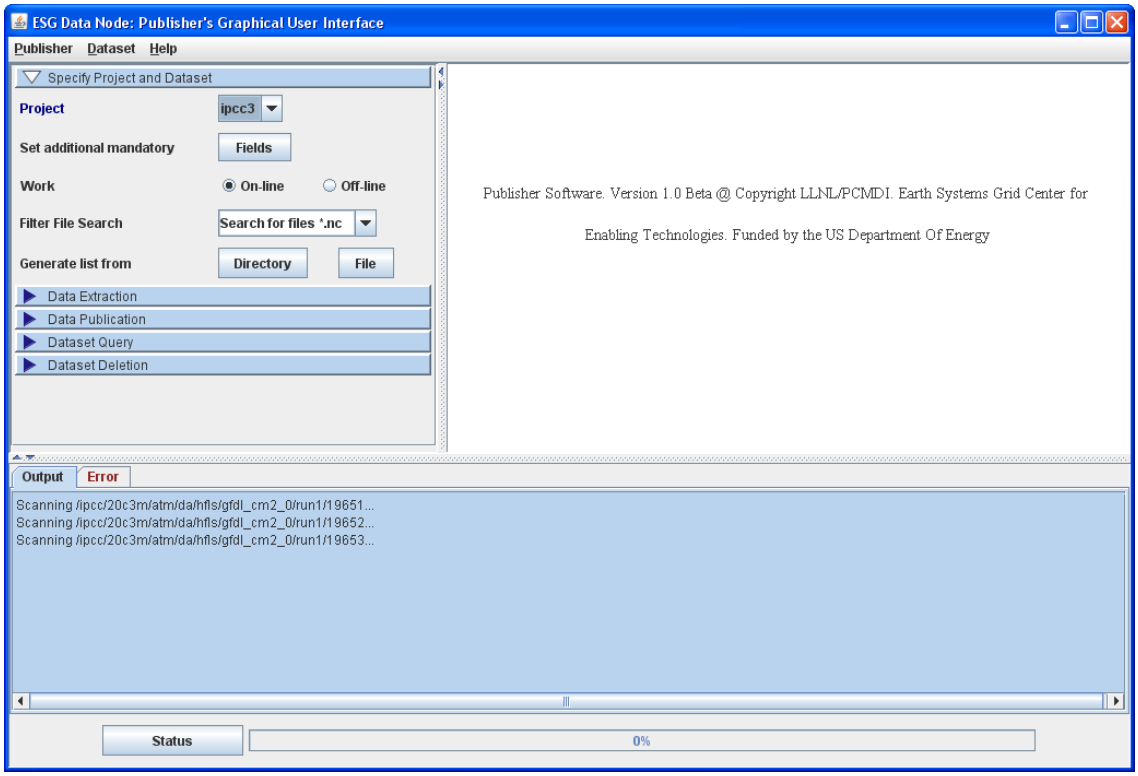

Figure 5: This is a screenshot of the new Java-based GUI publisher, which will allow data providers to remotely publish data to ESG gateways.

3.3 GridFTP for Replication and Download

ANL continues to enhance and support GridFTP as a high-performance download mechanism in ESGF. In the last six months, the ANL team has worked on the production readiness of the ESG infrastructure for distributing CMIP5 data. Specific tasks have included bug fixes to the GridFTP authorization library to support its deployment at DKRZ, and updates to the library to ensure compatibility with BADC's deployment. ANL has also provided GridFTP server testing and support, including for critical features that limit the root directory from which the server can access data, and for the use of the Java KeyStore framework on data nodes.

GridFTP solutions are also used for replication of core climate data across multiple sites. Many of the ESGF sites are being set up with the infrastructure needed for data replication.

3.4 Globus Online Integration

Globus Online is a data transfer management software-as-a-service provided by ANL and University of Chicago that builds on the GridFTP data transfer mechanism and provides reliable and high-performance data transfer solutions. Globus Online lets users "fire-and-forget" their transfer request and the service will manage the entire operation—monitoring performance, retrying failed transfers, recovering from faults automatically when possible, and reporting request status. Further, Globus Online auto-tunes the transfer to ensure best performance based on the transfer size and the number of files per transfer.

ESG can outsource data transfer management to Globus Online and focus on domain-specific tools and solutions for climate science research. Some of the concepts and solutions for the integration were demonstrated at the SC10 conference.

ANL is integrating the Globus Online transfer solution as an option for data downloads and replication in ESG. Globus Online is provided as one of the options in the ESG gateway for downloading selected data. The user experience includes user login, search and selection of data to download, and a series of UI-driven steps to submit the transfer request to Globus Online. The transfer uses the user's credentials from an ESG MyProxy server, and allows users to either download to their local machines or transfer to any other machine with a GridFTP server.

A prototype of the above workflow and some initial testing has been completed. Additional work is needed to improve the UI, configure ESGF OpenID providers as Globus Online-accepted identity providers, provide ESGF MyProxy servers as the login mechanism, and offer a cleaner workflow for first-time Globus Online registration. This work will be merged with the NCAR gateway codebase and be made available to the ESGF community in the next few months.

A similar effort to leverage Globus Online for replication use cases is underway. This work is focused on using the Secure Shell (SSH)-based client interface to Globus Online for automating the replication process by administrators. USC/ISI has enhanced the ESGF replication tool to support Globus Online as a transfer option, and ANL is providing documentation and support for testing and using this option. BADC has actively tested the replication tool, and ANL has provided enhanced some features for controlling performance parameters.

The above efforts have resulted in ANL developing a Java client library for the Globus Online Transfer API, which can be leveraged by other ESGF tools and services to integrate data transfer functionality. In collaboration with other ESGF partners, these features will be integrated with other search and download tools and CDAT in the coming months.

3.5 Federation Security

ANL has contributed to various federation security components, including:

1. Enhanced and maintained the ESGF trusted certificates repository and the white list of trusted services. Additional improvements include a Web page for downloading data and restructuring the stored archive to provide a better mechanism for access to the trust store.
2. Completed enhancements to the MyProxy client for ESGF end users to obtain a credential for accessing data, supporting a more streamline WebStart user experience. These enhancements allow the gateway to auto-configure various parameters at start-up, thus eliminating the need for users to provide this information, and allowing scripting for client-based access.
3. Worked with other team members on integrating ESGF security solutions to LAS, including providing input on solution architecture and performing hands-on installation and testing of the implemented solution on a LAS data server.
4. Designed a schema for representing federation-wide metadata regarding various services, including data access services as well as authorization and attribute services. ANL also has provided a Web form for ESGF node administrators for providing node information that can then be distributed to other nodes in the federation. Additional work is underway with other members

of the ESG-CET team to develop a federation-wide registry for consuming and disseminating this information.

3.6 Berkeley Storage Manager

The Berkeley Storage Manager (BeStMan) is an implementation of the Storage Resource Management (SRM) system used in ESG. BeStMan's architecture is based on a modular design that makes it easy to interface to various storage systems, including Mass Storage Systems (MSSs). Versions of BeStMan have been developed for High Performance Storage Systems (HPSS) at LBNL/NERSC, HPSS at ORNL, and MSS at NCAR. These versions of BeStMan handle different security mechanisms. They have been developed over the last two years, and in the last six months, they have been deployed for inclusion in the new ESGF system. The following tasks were accomplished during this reporting period:

- LBNL/NERSC, NCAR, and ORNL BeStMan servers were deployed and served ESG users in production through the new ESG gateway portal.
- BeStMan servers for the three sites mentioned above were used by the ESG Metadata Publishing tool for browsing all file information for files stored on deep storage to populate the metadata catalogues at the gateway.
- ORNL BeStMan has a new plug-in for two layer internal security. The ORNL ESG team developed a plug-in that combines BeStMan in the external security zone and the HPSS data-fetching program in the internal security zone through a secure communication channel. This security configuration was successfully deployed for next-generation ESGF systems at ORNL.
- Feature additions and bug fixes:
    - Fixed MSS request processing queue to resolve max queue size error.
    - Added MSS timeout for hanging MSS file retrieval requests to prevent overflowing of the MSS request queue. This feature times out by default when the MSS connections idle for more than 2 hours.
    - Added support for the Hypersonic 2 (H2) database in the BeStMan server, replacing BerkeleyDB for faster processing and smaller memory footprint.
    - Added support for fully qualified domain name (FQDN) caching from the resolved user identifiers (UIDs) for efficiency.
    - Added support for proxy host as defined by GLOBUS_HOSTNAME.
    - Fixed cached file processing during the restart. Prior to these changes, expired files were not properly cleaned up during the restart of the server, causing confusion when the files were requested later.
    - Fixed client error in SRM-copy during the srm_abort call to BeStMan server. Previously, waiting time between the status call and the client's abort request to the server caused confusion when the request failed.
- Client support was provided in the following areas:
    - Unauthorized access error to LBNL BeStMan server from NCAR data portal. The issue was resolved by using the correct hostname.

- Authentication errors to LBNL BeStMan server from NCAR data portal. This issue was resolved with proper data node entry in the grid-mapfile that was caused by a change of the client data node.
- Connection failure to ORNL BeStMan server. The expired grid host certificate caused connection failure. The issue was resolved by renewing the certificate.
- Connection failure to ORNL BeStMan server. The CA signing policy file was improperly formatted and had the wrong hostname for access_id_CA. The issue was resolved by properly entering the signing policy file.

3.7 DataMover-Lite

Based on user feedback, changes were made to DML requirements for authorizing the plug-in in GridFTP servers in the new ESGF system. The main updates were to support multiple ESGF gateways and http downloads. DML v3 has been fully integrated with the ESG authorization model and plug-ins. The stand-alone version (http://sdm.lblgov/dml/) and the WebStart version (http://datagrid.lbl.gov/dml3/jnlp/dml.html) have been used in production. The following tasks were accomplished during the reporting period:

- Updated documents on http://sdm.lblgov/dml/
- Updated DML WebStart on http://datagrid.lbl.gov/dml3/jnlp/dml.html
- Integrated WGet download script with DML downloads.
  - All http downloads from WGet script were integrated with DML WebStart version.
  - Integrated system works for both old and new WGet download scripts generated from ESG gateways.
  - WGet download scripts from ESG gateways can be selected as a single input file to DML. DML will parse the WGet scripts to download the files.
- Provided support for multiple ESG gateway MyProxy servers by developing a dropdown list that users can choose from.
  - Tested credential retrieval from MyProxy servers and GridFTP/http downloads.
- Bundled new ESGF CA certificates in the DML package.
- Collaborated with the gateway developmental team on developing a catalog-browsing feature within DML for catalog API support.
- Provided user support on local firewall issues.

3.8 Bulk Data Mover

BDM has been released and is being tested extensively with the ESG replication service. A few features were added to improve efficiency. The ESG-CET team has provided client support throughout the reporting period. These tasks include:

- Updated documents on http://sdm.lbl.gov/bdm/
- Added a command option—"-symlink"—for following soft links during the file transfers

- Provided support for proxy server for downloads with globus_hostname environment variable.
- Provided support for root directory browsing and file transfers.
- Provided support for the third-party GridFTP transfers.
- Added a command option—"-checkpoint"—for periodic summary updates.
- Provided support for concurrent source directory browsing to enhance performance.
- Provided support for concurrent threads for checking the existence of destination files.
- Provided support for source directory browsing with a non-recursive option.
- Provided support for directory transfers with a non-recursive option.
- Provided support for blank spaces in the file path.
- Added a command option—"-perf"— for network performance checking, including /dev/null and /dev/zero support.
- Provided support for dynamic connection adjustment.
- Tuned the back-end database for efficiency.
- Updated output summary display for convenience.
- Added graceful exit when the user proxy is invalid.
- Created BDM performance throughput plots for test runs from NERSC to ANU.
- Figure 6 shows approximately 5.6 gigabytes per second (Gbps) (695 megabytes per second) on average with approximately 6 Gbps at the peak.
    - https://sdm.lbl.gov/wiki/Software/BDM/BDMSamplePlots

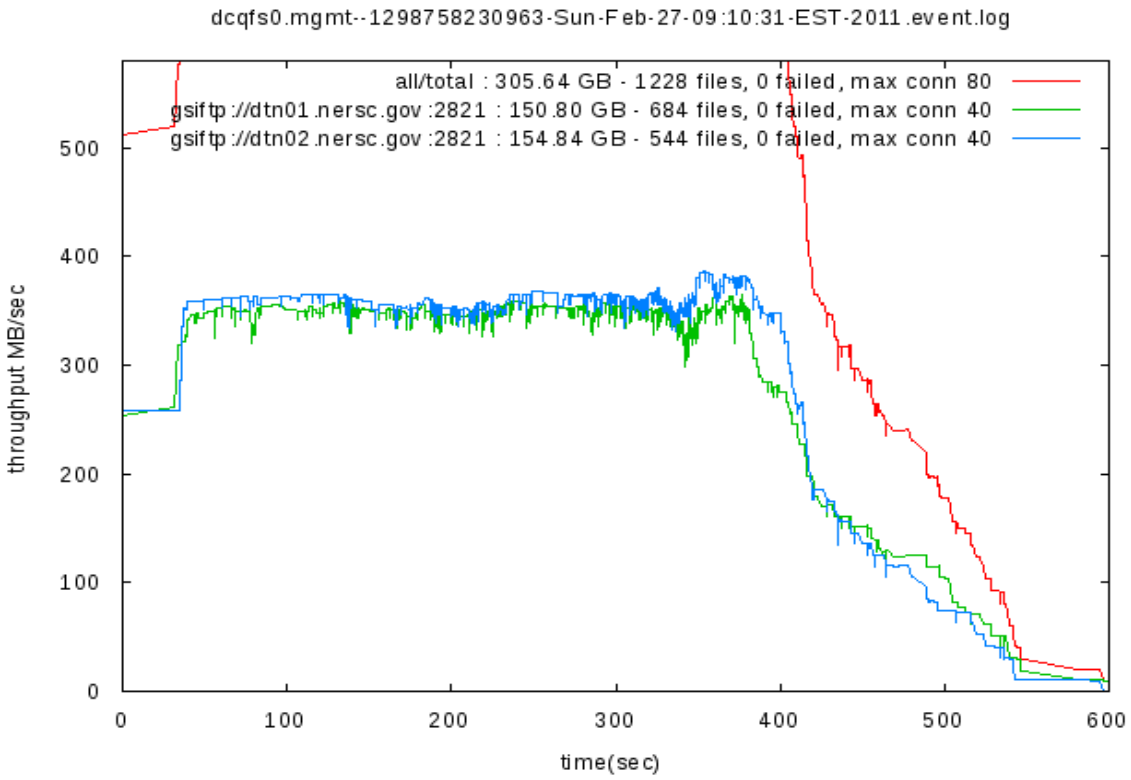

Figure 6: BDM transfer results shows throughput over time in seconds, while transferring 1,228 files in 305GB of climate data from two sources at LBNL/NERSC to one destination at ANU. This task was accomplished with 40 concurrency for each data source on a shared network with a 10-Gbps connection.

- Client support
    - BDM runs at DKRZ:
        - DKRZ directory structure was found to be coupled with several soft links. To enable soft link transfer, the "-symlink" option must be used in the BDM command. BDM will follow soft links to replicate files.
    - BDM runs for BADC:
        - Java API-based GridFTP clients (BDM and globus-url-copy) did not work with the cmip-bdm1.badc node, but C version of globus-url-copy did. It was suspected that the node had a gsiftp server certificate and not a host certificate. The issue was resolved by requesting a host certificate.
        - Bad certificate error caused by OpenJDK. (The error message was as follows: “Explanation: Authentication failed [Caused by: Miscellaneous failure. [Caused by: Bad certificate (java.security.SignatureException: SHA-1/RSA/PKCS#1: Not initialized)]]”.) JavaCog did not work with OpenJDK. IBM Java and other java versions affected this particular security exception even though the correct certification directory was used. The issue was resolved using the SUN/Oracle Java version.
    - GridFTP certificates issued at PCMDI.

- CA certificates (367b75c3 and 1149214e) and their signing policy files were missing in the PCMDI MyProxy trust store and ESG trusted certificates. Federated gsiftp server connections resulted in a connection error. The issue was resolved by updating the PCMDI server.
- PCMDI MyProxy trust store returned three signing policy files (159117b6, ae9c66bf, and 0084963c) with format errors. The issue was resolved by re-deploying the certificate tar file on the PCMDI node.

- BDM on ANU.
  - For BDM to run, globus_hostname had to be setup in the environment. When BDM ran on a machine with Dynamic Host Configuration Protocol (DHCP) enabled, the local IP address of the machine was incorrectly detected by Java. Most commonly, the detected IP address was the local loop-back address, 127.0.0.1. The issue was resolved by configuring BDM with the "--with-globus-hostname" option (e.g., --with-globus-hostname=64.34.58.128 or --with-globus-hostname=myhost.mydomain.tld).

3.9 Data Replication

The data replication team at ISI supported the expanded use and testing of the ESG replication client by deploying the software to two European ESGF partners' (BADC and DKRZ) sites. Personnel at these sites exercised and tested the features of the replication client and identified issues regarding its use function. The ISI team recorded the issues and tracked solutions in the ESGF bug tracker. Testing continued at PCMDI, USC/ISI and NCAR sites in the United States. DKRZ successfully used the ESG replication client to mirror data sets from BADC.

The replication client typically runs on a data node (the *mirror data node*) that will store a replica of an existing data set. The replication client takes a specified data set and begins querying the metadata catalog on a gateway to determine the list of files in that data set. The replication client also queries the gateway for a pointer to the THREDDS catalog on the data node (the *source data node*) where the data was originally published. Next, the replication client prepares data transfer requests to copy the data from the source data node to the mirror data node. These transfer requests may be prepared by the BDM client, the GridFTP globus-url-copy client, or the newly added Globus Online service. The replication client then copies the THREDDS catalog from the source data node to the mirror data node. After the data set has been successfully copied, the replication client invokes the publication API developed at LLNL to scan the copied files and create a THREDDS catalog for the replicated data set. Next, the replication client software compares the newly generated THREDDS catalog for the replicated data set to the THREDDS catalog from the source data node. The client then verifies that essential elements of the THREDDS catalog are consistent in both catalogs. If this check succeeds, then the replication client invokes the publication API to publish the replicated data set to a gateway. The replicated data set can then be discovered and accessed by ESG users.

Specific accomplishments during this reporting period for data replication include the following:

1. Packaged the software to make it easily deployable by participating sites.
2. Enhanced documentation to clarify the correct use of the software.
3. Enhanced the software in response to end-user needs.

a. Integrated a new command line switch with the ESG publication API.
b. Added Globus Online as an option agent for moving data.

4. Implemented bug fixes as needed.
5. Expanded the replication test plan as documented in the ESGF Wiki.

Over the next six-month reporting period, the ESG-CET team plans to provide continued support to its European partners to further deploy and test the replication client functionality; integrate the replication client source into the ESG-CET software package; hold follow-on discussions to determine additional requirements for replication; and implement changes as appropriate for the ESG and GO-ESSP collaborations.

3.10 Product Services

The PMEL team has completed work related to ESG goals on i) the LAS user interface client; ii) the Interactive Earth Science Data Visualization Gallery (vizGal) comparison Interface; iii) desktop access to OPeNDAP servers secured with the ESG authentication infrastructure; and iv) the connection between the ESG publisher generated THREDDS catalogs and LAS. The team also made improvements to the server-side analysis system (Ferret-THREDDS Data Server (F-TDS)) and completed the integration of LAS and F-TDS into the ESGF node installation script.

The PMEL team integrated the LAS User Interface Client with the ESG single-sign-on authentication, which enables the user to be successfully and automatically sent through the proper authentication process. The LAS Ajax services have also been improved so that any piece of the software infrastructure can generate links directly into LAS. As a result, the gateway interface, the publisher, and other software can direct users to an LAS interface. Single data sets or any sub-collections can be requested and displayed in the LAS User Interface client so it reveals only the data they have selected. Thus, user interaction and data discovery are kept in the gateway and users can focus on the selected data set. The team has also developed an Ensemble Axis prototype in the comparison interface so that data collections can be compared among ensemble members.

The PMEL team led the discussion and testing of direct OPeNDAP access from desktop UI clients to ESG data resources. The PMEL Ferret development team built the Ferret analysis application with beta netCDF software, which contained the necessary modifications (supplied by BADC) for using a temporary MyProxy certificate to access ESG data directly via OPeNDAP. Additionally, the team successfully tested accessing data from two sites: BADC and NASA/JPL.

PMEL worked in collaboration with the PCMDI team to include a LAS service element and access links to LAS in the THREDDS catalog metadata for data published in ESG. Now, when data are published with the generation of LAS metadata turned on, these links are automatically active in the TDS html interface and can also be harvested from the catalog for other uses. These LAS linking and configuration facilities depend on the work completed with PCMDI to ensure all of the necessary metadata for LAS configuration is moved out of netCDF data sources and into catalog metadata during the publication process.

The PMEL team also undertook work on the F-TDS so that it works with all netCDF data types including character variables. The improved F-TDS preserves all of the existing variable metadata and variable names when serving data. Additional accomplishments include working with PCMDI to automate the installation and configuration of LAS and F-TDS using the ESGF data node installation

script. These two services are installed, configured, and started automatically when a node site installs the node software stack using the installation scrip.

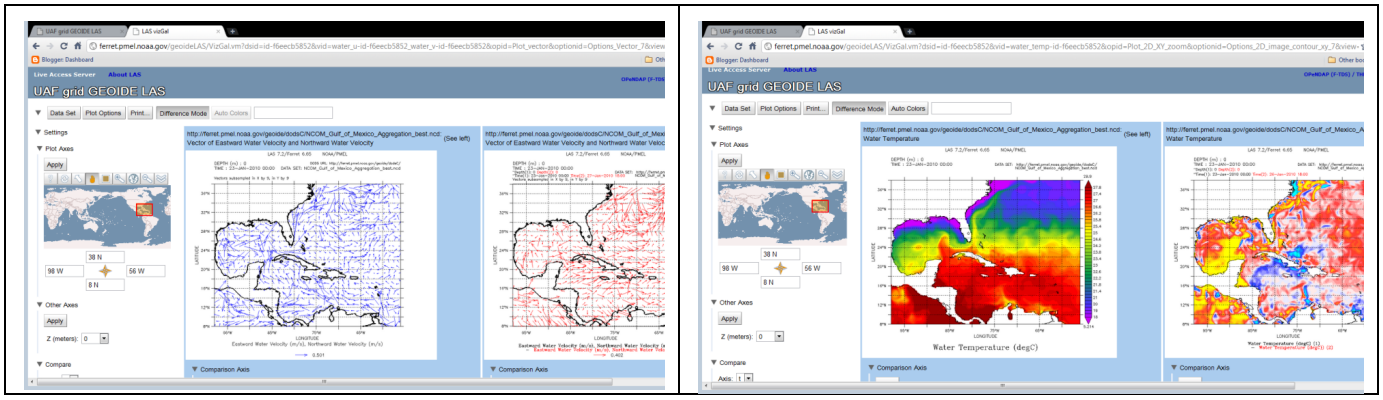

Figure 7: ESG's project goals are to allow easy intercomparison of scalar and vector data using a Web browser, facility data exploration, and data discovery within the enterprise system. State-of-the-art UI tools, which leverage existing and new server-side capabilities in the ESG product services infrastructure (LAS), were used to build a rich and interactive UI that extends the comparison interface to vector and scalar data.

4 ESG-CET Group Meetings

The ESG-CET executive committee holds weekly conference calls on Tuesdays at 9:00 a.m. Pacific Daylight Time (PDT). At these meetings, the committee discusses the priorities and issues that will make up the agenda for the weekly project meetings, which are held on Thursdays at 12:00 p.m. PDT via the Access Grid (AG). At the weekly project meetings, the entire team discusses project goals, design and development issues, technologies, timelines, and milestones. Given the need for more in-depth conversation and examination of work requirements, several face-to-face meetings are planned for the next reporting period.

5 Collaborations

To effectively build an infrastructure that can accommodate petascale data management and analysis, the ESG-CET team established connections with researchers/scientists involved in other DOE Office of Science SciDAC-funded projects and programs. In particular, collaborations were developed with the following groups: TeraGrid Science Gateways, Earth System Curator, NOAA's GIP, METAFOR, World Meteorological Organization (WMO) Information System, Scientific Computing and Imaging Institute at the University of Utah, SciDAC Visualization and Analytic Center for Enabling Technologies (VACET), SciDAC Scientific Data Management Center (SDM), SciDAC Center for Enabling Distributed Petascale Science (CEDPS), Southern California Earthquake Center, Tech-X Corp., NASA Langley, NASA Goddard, GO-ESSP, and many others.

5.1 2011 Global Organization for Earth System Science Portal Workshop

Dean N. Williams, Steve Hankin, and Don Middleton are members of the GO-ESSP steering committee responsible for organizing the 2011 workshop, which will be held in Asheville, North Carolina, USA, and hosted by the National Climatic Data Center (NCDC).

The GO-ESSP workshop facilitates the organization and implementation of an infrastructure for full data sharing among a consortium spanning continents, countries, and intergovernmental agencies. This consortium envisions an environment that allows users open access to petabytes of model-generated, satellite, and in situ data including physical, biogeochemical, and ecosystem content.

5.2 LLNL Computation Directorate External Review

Dean N. Williams presented ESG-CET at the LLNL Computation Directorate External Review. The presentation entitled, "Enabling Worldwide Access to Petascale Climate Data" reported on ESG and discussed outstanding issues, future needs, and scientific directions for the climate community. The talk highlighted ESG as the cornerstone for distributing climate data to the community.

5.3 DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization

Dean N. Williams and Gary Strand attended the DOE ASCR (Advanced Scientific Computing Research) 2011 Workshop on Exascale Data Management, Analysis, and Visualization held in Houston, Texas. The two-day workshop focused on scientific data analysis and visualization for charting the research roadmap necessary to support scientific discovery on exascale systems. Strand presented information on the needs of the climate community, which included the need for ESG.

5.4 Terabit Networks for Extreme Scale Science

Dean N. Williams, Ian Foster, Alex Sim, and Galen Shipman represented ESG-CET at the DOE ASCR Terabit Networks for Extreme Scale Science. The workshop, held in Rockville, Maryland, focused on advanced network services and middleware that integrate users and science applications to the underlying terabit network infrastructure. Williams gave a talk explaining ESG and how its served applications are data and network intensive. He also suggested that if the network can provide some new functionalities and services, it might be able to serve a broader range of disciplines.

5.5 American Meteorological Society Python Short Course

Dean N. Williams presented ESG-related topics at the American Meteorological Society conference in a two-day short course symposium entitled, "Using Python in Climate and Meteorology: Advanced Methods." The short course, held in January in Seattle, Washington, was designed in modules covering various applications for which Python is used, such as Input/Output, visualization, diagnosis, languages coupling, etc.). The course included projects such as CDAT and ESG, showing how Python is used to manipulate and access petabyte data sets.

5.6 National Climate Assessment Teleconference

Dean N. Williams presented for ESG-CET at the National Climate Assessment (NCA) Teleconference held December 13, 2010. NCA aims to incorporate advances in the understanding of climate science into larger social, ecological, and policy systems, with the goal of providing integrated analyses of impacts and vulnerability. ESG was discussed as a possible infrastructure for their effort.

5.6.1 DAARWG Workshop

Dean N. Williams attended the NOAA Advisory's Board Data Archive and Access Requirements Working Group (DAARWG) in early December 2010—an important working group of NOAA's Science Advisory Board (SAB). DAARWG evaluates data archiving and access requirements from all NOAA observing systems and computational models, as well as from relevant non-NOAA sources. Its charter is to provide scientific advice and broad direction regarding the wide range of data, information, and products that NOAA archives, and ways in which the agency can best provide access to these resources. ESG was mentioned throughout the workshop as part of their end-to-end data environmental data management lifecycle.

5.7 NOAA Global Interoperability Program

NOAA's GIP promotes coordination of software infrastructure development across agencies, the weather and climate communities, modeling and data services, and research and operational centers.

- The Curator project, now hosted under GIP, continues to be a highly productive collaboration with ESG-CET. Curator serves as an active liaison between the EU-based METAFOR project, which is developing a CIM for CMIP5/IPCC, and ESG-CET, which will be delivering data to the global community. Through this effort, the ESG-CET team has continued to develop the ESG system's ability to harvest CIM output from METAFOR and represent the results using the "model trackback" capabilities. These capabilities effectively combine the METAFOR questionnaire with the query and browse functions of ESG.
- The "Standardized Analysis Workflows for Climate Models" project continues under GIP. The project addresses the need for Web-accessible tools via ESG that are suitable for the analysis and intercomparison of climate model outputs. Special focus has been on collections of model outputs that comprise CMIP5 multi-model ensembles—models that do not necessarily share the same grid coordinate structure and may include multi-tile "gridspec" coordinate systems.

5.8 National Science Foundation TeraGrid Science Gateways Program

ESG-CET, NOAA GIP (Curator), and Purdue University staff collaborated on a proposal to develop a prototype Environmental Science Gateway and were granted support from TeraGrid's Science Gateway program for two years. The basic objectives of this project include developing a gateway-based model execution capability, a supporting workflow system that can operate on the TeraGrid gateway, and data publishing and harvesting capabilities that can interoperate with an integrated rule-oriented data system–(iRODS-) based archive as well as disk-hosted model output. During this reporting period, the ESG-CET team made substantial progress developing an Atom-based ingests capability, new portal job submission capabilities, and provisions for Earth System Modeling Framework (ESMF)-based models to produce CIM-compliant metadata for ingestion into the ESG gateway.

5.9 The North American Regional Climate Change Prediction Program

Since its inception, the NARCCAP program has been using ESG infrastructure to successfully publish, manage, and share regional climate model data with the broader climate science community. NARCCAP staff continues to publish new data sets into ESG, even as NARCCAP heads into its final project year. Middleton attended the annual NARCCAP Principal Investigator (PI) meeting in 2010 at NCAR in Boulder, Colorado.

6 Outreach, Papers, Presentations, and Posters

This section describes ESG-CET outreach activities for this reporting period, which include published papers and presented talks.

6.1 Papers and Posters

- "***Finding Tropical Cyclones on a Cloud Computing Cluster: Using Parallel Virtualization for Large-Scale Climate Simulation Analysis,*"** D. Hasenkamp, **A. Sim**, M. Wehner, K. Wu, Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010.
- "***Finding Tropical Cyclones on Clouds,*"** D. Hasenkamp, under the supervision of **A. Sim**, M. Wehner, K. Wu, ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10), 2010.
- ***"Adaptive Transfer Adjustment in Efficient Bulk Data Transfer Management for Climate Datasets,"* A. Sim**, M. Balman, **D. Williams, A. Shoshani, V. Natarajan**, Proceedings of the 22nd International Conference on Parallel and Distributed Computing and Systems (PDCS 2010), 2010.
- ***"The Interactive Earth Science Data Visualization Gallery (vizGal),"* Roland Schweitzer**, Weathertop Consulting, LLC, College Station, TX; and **S. Hankin**, J. Malczyk, A. Manke, and K. M. O'Brien, 27th Conference on Interactive Information Processing Systems (IIPS), AMS Annual Meeting, Seattle, Washington (2011).

6.2 Talks

- **"*ESG and Cloud Computing with an Experience from Exploring Cloud for Parallelizing Tropical Storm Tracking,*" A. Sim**, Expedition Workshop, Seeing Through the Clouds: Exploring Early Communities and Markets Streamlined by Open Government Principles, October 19, 2010.

6.3 NARCCAP PI Meeting

Don Middleton attended the October 19, 2010, Fall NARCCAP PI meeting help at NCAR in Boulder Colorado, and provided a brief update on ESG status.

6.4 Dynamic, Distributed, Data-Intensive Programming Abstractions and Systems

Don Middleton presented "Cyberinfrastructure and the Global Environmental Data Challenge" at this event, also known as "3DPAS," in February 2011 in Edinburgh, Scotland, UK. 3DPAS is a joint venture of the US National Science Foundation, the UK e-Science Institute, and UK government research programs.

6.5 NASA-IPCC Data System Workshop

ESG team members presented multiple talks at the NASA-IPCC Data System Workshop held at NASA Goddard Space Flight Center on November 9–10, 2010. The purpose of this workshop was to discuss a NASA technical implementation plan to share observations for CMIP5 and create an effective

partnering relationship between NASA and other federal agencies. This meeting followed a successful workshop at PCMDI on October 12–13, 2010, that identified potential observational data sets to be shared for the study.

- Dean N. Williams, Luca Cinquini, Dan Crichton, and Galen Shipman presented "Partnership for 21st Century Earth System Science."
- Luca Cinquini, Amy Braverman, Dan Crichton, Chris Mattmann, and Dean Williams presented "Requirements and Process for Publishing Observational Datasets into the global CMIP5 Archive."
- Gary Strand presented "CESM, CMIP5, ESG, and Data Management."
- Galen M. Shipman, Sig Christensen, Luca Cinquini, John Harney, Misha Krassovski, Ross Miller, Giri Palanisamy, and Feiyi Wang presented "ESG – Observation Data Set Integration."

6.5.1 PCMDI – NASA Meeting

Dean N. Williams, Luca Cinquini, Gavin Bell, Daniel Crichton, Bob Drach represented ESG-CET at PCMDI—a NASA meeting sponsored by NASA and DOE. The meeting was held at LLNL in Livermore, California on October 12-13, 2010. As part of an ongoing effort to get NASA and NOAA observational data into the CMIP5 archive, this group is looking to include observational data sets to help score models when evaluating ensemble predictions and to improve the understanding of physical processes.

6.5.2 Greenhouse-Gas Information Systems

Dean N. Williams represented ESG at the Greenhouse-Gas Information Systems requirement meeting held at LLNL in early October 2010. Continued discussions are planned for the 2011 fiscal year regarding the potential role of ESG within the proposed mission operation center.

6.6 Posters

No posters presented for the reporting period.